Amendments to the Specification:

Please replace the paragraph at page 2, lines 9-16 with the following amended paragraph:

As previously mentioned, each DNA molecule contains many genes. A gene is a specific sequence of nucleotide bases. These sequences carry the information required for constructing proteins. A protein is a large molecule formed of one or more chains of amino acids in a specific order. Order is determined by base sequence of nucleotides in the gene coding for the protein. Each protein has a unique function. ~~well-defined functionality. A DNA sequence consists of many biologically distinct regions. For the purpose of this application, Applicants distinguish between intergenic DNA and genes. Many of the genes in mammalian cells are "split genes". A split gene consists of coding and non-coding sequences. The coding sequences in a gene are contained within exonic regions (exons), that appear sequentially separated by long regions referred to as introns.~~ In a DNA molecule, there are protein-coding sequences (genes) called "exons", and non-coding-function sequences called "introns" interspersed within many genes. The balance of DNA sequences in the genome are other non-coding regions or intergenic regions.

Please replace the paragraph at page 3, lines 3-25 with the following amended paragraph:

Gene identification and gene discovery in newly sequenced genomic sequences is one of the most timely computational questions addressed by bioinformatics scientists. Popular gene finding systems include Glimmer, Geumark, Genscan, Genie, GENEWISE, and Grail (See Burge, C. and S. Karlin, "Prediction of complete gene structures in human genomic DNA," J Mol. Biol., 268:78-94, 1997; Salzberg, S. et al., "Microbial gene identification using interpolated Markov models," Nucl. Acids Res., 26(2):544-548, 1998; Xu, Y. et al., "Grail: A multi-agent neural network system for gene identification," Proc. of the IEEE, 84(10):1544-1552, 1996; Kulp, D. et al., "A generalized hidden Markov model for the recognition of human genes in DNA," in ISMB-96: Proc. Fourth Intl.

Conf. Intelligent Systems for Molecular Biology, pp. 134-141, Menlo Park, Calif., 1996, AAAI Press; Borodovsky, M. and J.D. McIninch, "Genemark: Parallel gene recognition for both DNA strands," Computers & Chemistry, 17(2):123-133, 1993; and Salzberg, S. et al. eds., Computational Methods in Molecular Biology, Vol. 32 of New Comprehensive Biochemistry, Elsevier Science B.V., Amsterdam, 1998). The annotations produced by gene finding systems have been made available to the public. Such projects include the genomes of over thirty microbial organisms, as well as Malaria, Drosophila, C.elegans, mouse, Human chromosome 22 and others. For instance, Glimmer has been widely used in the analysis of many microbial genomes and has reported over 98% accuracy in prediction accuracy (See Fraser, C. M. et al., "Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi," Nature 390(6660):580-586, December 1997). Genie (D. Kulp et al. above) has been deployed in the analysis of the Drosophila genome, and Genscan (C. Burge and S. Karlin above) was used for analysis of human chromosome 22.

Please replace the paragraph at page 4, lines 13-23 with the following amended paragraph:

On a very high level, genes in human DNA and many other organisms have a relatively simple structure. All eukaryotic genes, including human genes, are thought to share a similar layout. This layout adheres to the following "grammar" or pattern: start codon, exon, ~~(intron-exon)^n~~ (intron-exon)$_n$, stop codon. The start codon is a specific 3-base sequence (e.g. ATG) which signals the beginning of the gene. Exons are the actual genetic material that code for proteins as mentioned above. Introns are the spacer segments of DNA whose function is not clearly understood. And finally stop codons (e.g. TAA) which signal the end of the gene. The notation (intron-exon)$_n$, simply means that there are n alternating intron-exon segments. Genes identification procedures has to take into account other important issues such as polyA tail, promoters, pseudo-genes, alternative splicing and other features.

Please replace the paragraph at page 7, lines 3-9 with the following amended paragraph:

By way of background, ~~ligated~~ exons are the sequence regions that are translated into proteins ~~form a sequence that is translated into a protein~~ by a simple but still computationally mysterious mechanism of splicing that takes place after the DNA sequence has been transcribed into RNA. The process starts by spliceome proteins that recognize the splice signals, followed by a step where the introns are cut out (spliced out), and ending in a phase where the consecutive exons are "glued" together into a single sequence that is translated into a protein. Intuitively speaking this process is performed on an RNA "image" of the genomic sequence.

Please replace the paragraph at page 8, lines 1-5 with the following amended paragraph:

The present invention is a system for the combination of individual experts which is learned from data. Unlike the prior art, such a system exploits learned dependencies between experts and forms a prediction maximally consistent with known gene data. Statistically, predictions of the invention system will then have the potential to generalize to genes undiscovered by any of the individual experts ~~refine the boundaries and verify the predictions made by experts~~.

Please replace the paragraph at page 10, lines 7-18 with the following amended paragraph:

Bayesian network probabilistic models provide a flexible and powerful framework for statistical inference as well as learning of model parameters from data. The goal of inference is to find a distribution of ~~one or more~~ a random variable in the network conditioned on evidence (known values) of other variables. Bayesian networks encompass efficient inference algorithms, such as Jensen's junction tree (Jensen, F.V., An Introduction to Bayesian Networks, Spring-Verlag, 1995) or Pearl's message passing (Pearl, J., Probabilistic reasoning in intelligent systems, Morgan Kaufmann, San Mateo, Calif. 1998).

Inside a learning loop, such algorithms may be used to efficiently estimate optimal values of a model's parameters from data (for instance, see Jordan, M.I. ed., Learning in Graphical Models, Kluwer Academic Publishers, 1998). Furthermore, techniques exist that can optimally determine the topology of a Bayesian network together with its parameters directly from data.

Please replace the paragraph at page 11, lines 9-18 with the following amended paragraph:

Gene combiner parameters, probability tables $P(E_1|Y)$ and $P(Y)$, are learned from a training dataset of nucleotide sequences by statistically calculating $P(E_1|Y)$ and $P(Y)$ of all individual predictors E_1 and labeled for ground truth Y. For instance, a maximum likelihood (ML) estimate of these parameters for a training set of N nucleotides is

$$P(E_i = e \mid Y = y) = \frac{\#\, E_i = e, Y = y}{N}$$

where e denotes the prediction of an expert system i, e ε {intron, exon}, and y is the combined prediction, y ε {intron, exon}. # E_i = e, Y = y denotes the number of cases in the training dataset where the prediction of expert system i is e and the ground truth ~~combined prediction~~ is y. Alternative estimates of these parameters may be obtained using MAP (maximum a posteriori) estimation.

Please replace the paragraph at page 14, lines 6-13 with the following amended paragraph:

For that purpose, Applicants assumed that each individual expert system provides the following binary decision. An expert system produces a single labeling for every nucleotide in a sequence: E if the nucleotide is a part of an exon and I if it belongs to an intron ~~or an intergenic region~~. Using the notation of Applicants' models, E_i ε {E,I} for an expert i. Similarly, a combined decision Y is either E or I. Parameters of each of the four above-discussed models of Bayesian network combiners 28, 31, 40, 51 were learned using a standard

maximum likelihood estimation in the Bayesian network framework. All prediction results were then obtained using a five-fold cross-validation.

Please replace the paragraph at page 14, lines 7-12 with the following amended paragraph:

An exon is said to be exactly predicted 47 only if both its ending and beginning points coincides with that of a true exon. An exon is said to be missed 57 if there is no overlap with any of the predicted exons. ME gives the percentage of missed exons 57 whereas WE gives the percentage of wrongly or overpredicted exons 49. To compute these two numbers (ME and WE), Applicants look for any overlap between a true and a predicted exon. ~~Wrong exon (WE) prediction implies the prediction has no overlap with a true exon.~~